Interpretations and generalizations of Kendall's rank correlation test

Terpstra, Teunis Jannes

SUMMARY

The starting point of all tests developed in this thesis is KENDALL's original rank correlation test. The usual interpretation of this test is the following.

By means of k independent observations (x_h, y_h), $h \leqslant k$ from a pair of variables $(\underline{x}, \underline{y})$, we want to test the hypothesis H of independence of the variables $\underline{x}$ and $\underline{y}$ against the alternative hypothesis H_1 of a positive *rank correlation* of $\underline{x}$ and $\underline{y}$.

The critical region for this test is defined by means of KENDALL's so-called rank correlation statistic $\underline{S}$ (a brief explanation of the principles of the theory of NEYMAN and PEARSON for testing statistical hypotheses and some properties of the underlying probability theory are given in Chapter 1).

Now another interpretation can be given to the test by considering the observations $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$ as single and completely independent observations of two sets of variables $\underline{x}_1, \ldots, \underline{x}_k$ and $\underline{y}_1, \ldots, \underline{y}_k$, which are asummed to be *concordant*.

Two sets of variables $\underline{x}_1, \ldots, \underline{x}_k$ and $\underline{y}_1, \ldots, \underline{y}_k$ will be called concordant if non-homogeneity of the sets can only occur simultaneously and in such a way that (roughly speaking) the rank orders of both sets of variables are 'on the average' the same.

Then KENDALL's rank correlation statistic may also be used for testing the hypothesis H of homogeneity of two concordant sets of variables $\underline{x}_1, \ldots, \underline{x}_k$ and $\underline{y}_1, \ldots, \underline{y}_k$.

Both tests can be generalized.

The first, by considering a set of variables $(\underline{x}^{(1)}, \ldots, \underline{x}^{(m)})$, $m \geqslant 2$, from which k independent observations $(x_h^{(1)}, \ldots, x_h^{(m)})$, $h \leqslant k$ are taken. It is allowable for single observations $x_h^{(\alpha)}$, $h \leqslant k$, $\alpha \leqslant m$ to be missing. By means of these observations we want to test the hypothesis H of independence of the variables $\underline{x}^{(1)}, \ldots, \underline{x}^{(m)}$.

The second, by considering any number m, $m \geqslant 2$, of concordant sets of variables $\underline{x}_1^{(\alpha)}, \ldots, \underline{x}_k^{(\alpha)}$, $\alpha \leqslant m$. Further not only one, but any number of observations may be taken from each of the variables $\underline{x}_h^{(\alpha)}$, where it is also allowable that no observation is taken. All observations are assumed to be completely independent. By means of these observations we want to test the hypothesis H of homogeneity of each set of variables $\underline{x}_1^{(\alpha)}, \ldots, \underline{x}_k^{(\alpha)}$.

For these problems generalized rank correlation statistics $\underline{S}$ and $\underline{Z}$ are defined, which are, for the special case of two sets of observations $x_1,\ldots,x_k$ and $y_1,\ldots,y_k$, identical with (respectively equivalent to) KENDALL's original rank correlation statistic.

If in the second problem with $m = 2$ it is assumed that $\underline{y}_1 < \underline{y}_2 < \ldots < \underline{y}_k$ with probability one, then concordancy of the two sets of variables $\underline{x}_1,\ldots,\underline{x}_k$ and $\underline{y}_1,\ldots,\underline{y}_k$ means an upward *trend* of the first set of variables. Thus two statistics $\underline{T}$ and $\underline{W}$ follow from the generalized rank correlation statistic $\underline{S}$, which are appropriate for testing against trend.

All tests developed belong to a class C of conditional tests, which is defined in Chapter 1, section 1.10. For this class of tests two general theorems on the respective consistency and non-consistency of the tests are shown (cf. the last chapter). From these general theorems conditions for consistency and non-consistency of the special tests developed, are obtained.

Throughout the treatment of the tests it is assumed that the distribution-functions of the variables considered may possess discontinuities, so that equal observations (ties) may occur with a positive probability.

SAMENVATTING

Uitgangspunt van alle toetsen, welke beschreven worden, is KENDALL's rangcorrelatie-toets, waaraan gewoonlijk de volgende interpretatie wordt gegeven.
Door middel van k onderling onafhankelijke waarnemingen (x_h, y_h), $h \leqslant k$ van een paar variabelen $(\underline{x}, \underline{y})$, willen we de hypothese H, inhoudende dat $\underline{x}$ en $\underline{y}$ onderling onafhankelijk zijn, toetsen tegen de alternatieve hypothese H_1, inhoudende dat $\underline{x}$ en $\underline{y}$ een positieve *rangcorrelatie* bezitten.
De kritieke zone van de toets wordt gedefiniëerd met behulp van KENDALL's zo genaamde rangcorrelatie-grootheid $\underline{S}$.

Het is echter ook mogelijk een andere interpretatie aan de toets te geven. We beschouwen de waarnemingen $x_1, \ldots, x_k$ en $y_1, \ldots, y_k$ dan als volledig onafhankelijke en enkele waarnemingen van twee overeenstemmende rijen van variabelen $\underline{x}_1, \ldots, \underline{x}_k$ en $\underline{y}_1, \ldots, \underline{y}_k$. Hierbij worden twee rijen van variabelen *overeenstemmend* genoemd, indien ze slechts gelijktijdig niet gelijk verdeeld kunnen zijn en wel op zodanige wijze, dat (globaal gezegd) gemiddeld eenzelfde rangschikking naar opklimmende grootte bestaat.
Dan kan de rangcorrelatie-grootheid $\underline{S}$ gebruikt worden voor het toetsen van de hypothese H, inhoudende dat elk der rijen van variabelen gelijk verdeeld is.

Beide problemen kunnen worden gegeneraliseerd.

Het eerste door in plaats van twee, een groep van m ($m \geqslant 2$) variabelen $(\underline{x}^{(1)}, \ldots, \underline{x}^{(m)})$ te beschouwen, waarvan k onderling onafhankelijke waarnemingen $(x_h^{(1)}, \ldots, x_h^{(m)})$, $h \leqslant k$ worden genomen. Hierbij wordt toegelaten dat waarnemingen $x_h^{(\alpha)}$, $h \leqslant k$, $\alpha \leqslant m$ ontbreken. Met behulp van deze waarnemingen willen we de hypothese H toetsen, inhoudende dat de variabelen $\underline{x}^{(1)}, \ldots, \underline{x}^{(m)}$ volledig onafhankelijk zijn.

Het tweede probleem kan worden gegeneraliseerd door in plaats van twee overeenstemmende rijen van variabelen, m ($m \geqslant 2$) overeenstemmende rijen van variabelen $\underline{x}_1^{(\alpha)}, \ldots, \underline{x}_k^{(\alpha)}$, $\alpha \leqslant m$ te beschouwen. Van elk der variabelen $\underline{x}_h^{(\alpha)}$, $h \leqslant k$, $\alpha \leqslant m$ wordt bovendien niet slechts één waarneming genomen, maar een willekeurig aantal, waarbij ook wordt toegelaten, dat waarnemingen ontbreken. Verondersteld wordt dat alle waarnemingen volledig onafhankelijk zijn. Dan willen we de hypothese H toetsen, inhoudende dat voor iedere $\alpha \leqslant m$ de variabelen $\underline{x}_1^{(\alpha)}, \ldots, \underline{x}_k^{(\alpha)}$ eenzelfde waarschijnlijkheidsverdeling bezitten.

Voor deze problemen worden toetsingsgrootheden $\underline{S}$ en $\underline{Z}$ gedefiniëerd, welke voor het speciale geval van twee rijen van waarnemingen $x_1,\ldots,x_k$ en $y_1,\ldots,y_k$ identiek zijn met (respectievelijk gelijkwaardig zijn met) KENDALL's oorspronkelijke rangcorrelatie-grootheid $\underline{S}$.

Onderstellen we dat voor het tweede probleem, waarbij $m = 2$ wordt genomen, $\underline{y}_1 < \underline{y}_2 < \ldots < \underline{y}_k$ is met waarschijnlijkheid 1, dan komt overeenstemming van de twee rijen variabelen $\underline{x}_1,\ldots,\underline{x}_k$ en $\underline{y}_1,\ldots,\underline{y}_k$ overeen met een stijgend *verloop* van de rij $\underline{x}_1,\ldots,\underline{x}_k$. Op deze wijze volgen uit de gegeneraliseerde rangcorrelatiegrootheid twee toetsingsgrootheden $\underline{T}$ en $\underline{W}$, welke geschikt zijn voor het toetsen tegen verloop.

Alle hierboven genoemde toetsen behoren tot een klasse C van voorwaardelijke toetsen, welke gedefiniëerd wordt in hoofdstuk 1, paragraaf 1.10. In het laatste hoofdstuk worden voor deze klasse van toetsen twee theorema's bewezen met betrekking tot de asymptotische doeltreffendheid der toetsen. Met behulp hiervan worden theorema's verkregen betreffende de asymptotische doeltreffendheid der speciale toetsen, welke zijn beschreven.

Bij de behandeling der toetsen beperken we ons niet tot variabelen, die slechts continue verdelingsfuncties bezitten, zodat gelijke waarnemingen met een positieve waarschijnlijkheid kunnen voorkomen.